American Educational Research Journal
April 2017, Vol. 54, No. 2, pp. 241–278
DOI: 10.3102/0002831216687531
© 2017 AERA. http://aerj.aera.net

The Impact of Performance Ratings on Job Satisfaction for Public School Teachers

Cory Koedel
Jiaxi Li
University of Missouri
Matthew G. Springer
Vanderbilt University
Li Tan
University of Missouri

Spurred by the federal Race to the Top competition, the state of Tennessee implemented a comprehensive statewide educator evaluation system in 2011. The new system is designed to increase the rigor of evaluations and better differentiate teachers based on performance. The use of more differentiated ratings represents a significant shift in education policy. We merge teacher performance evaluations from the new system with data from post-evaluation teacher surveys to examine the effects of the differentiated ratings on job satisfaction for teachers. Using a regression-discontinuity design, we show that higher ratings under the new system causally improve teachers' perceptions of work relative to lower ratings. Our findings offer the first causal evidence of which we are aware on the relationship between performance ratings and job satisfaction for individual teachers.

Cory Koedel is an associate professor of economics and public policy at the University of Missouri-Columbia, Department of Economics and Truman School of Public Affairs. His research focuses on teacher quality and compensation, the effects of curriculum materials on student achievement, and the efficacy of higher education institutions.

Jiaxi Li received his PhD in economics from the University of Missouri-Columbia in 2014. His research focuses on teacher evaluation policies.

Matthew G. Springer is an assistant professor of public policy and education at Vanderbilt University's Peabody College, 230 Appleton Place, Nashville, TN, 37204. USA; email: matthew.springer@vanderbilt.edu. His research focuses on incentives, accountability, and compensation.

Li Tan is a PhD candidate in economics at the University of Missouri-Columbia. His research focuses on retirement policy and teacher evaluation.

KEYWORDS: personnel evaluation, teacher evaluation, job satisfaction, teacher quality, regression discontinuity

Introduction

State and local education agencies across the United States are working to improve teacher quality through the adoption of rigorous teacher evaluation systems.[1] The development of these systems is motivated in part by a large body of research showing that teachers differ dramatically in their effectiveness (for a recent overview, see Hanushek & Rivkin, 2010). Beyond directly affecting students' short-term academic success, more effective teachers have also been shown to positively affect later-life outcomes for students, including wages, college attendance, and teenage childbearing (Chetty, Friedman, & Rockoff, 2014). The variation in teacher quality that has been consistently documented in empirical research is not reflected in the traditional teacher evaluation systems that are used in most school districts in the United States, which tend to assign similar and favorable performance ratings to the overwhelming majority of teachers (Kraft & Gilmour, 2016; Weisberg, Sexton, Mulhern, & Keeling, 2009). A rationale for developing new, more rigorous evaluation systems is to construct ratings that better reflect empirically observed performance differences across teachers, which can subsequently be used to improve workforce quality in a number of ways (Boyd, Lankford, Loeb, & Wyckoff, 2011; Condie, Lefgren, & Sims, 2014; Dee & Wyckoff, 2015; Ehlert, Koedel, Parsons, & Podgurksy, 2014; Goldhaber, Cowan, & Walch, 2013; Goldhaber & Theobald, 2013; Hanushek, 2009; Papay, Taylor, Tyler, & Laski, 2016; Springer, Swain, & Rodriguez, 2016; Steinberg & Sartain, 2015; Taylor & Tyler, 2012; Winters & Cowen, 2013).

In many respects, Tennessee has become a national leader for the design and implementation of teacher evaluation systems.[2] In January 2010, the Tennessee General Assembly passed Senate Bill 5, also known as the First to the Top (FTTT) Act, reforming dozens of areas of state education policy. Most notably, the new law requires that all school personnel be evaluated annually, and personnel decisions must be based at least in part on the evaluations. The four models approved by the Tennessee State Board of Education to evaluate teachers each contain several components that are combined to produce measures of overall teacher effectiveness. Unlike the traditional teacher evaluation systems examined by Weisberg et al. (2009) and recently revisited by Kraft and Gilmour (2016), the new ratings in Tennessee are sufficiently dispersed to indicate meaningful variation in effectiveness across the teaching workforce.

We examine one aspect of how the new, more rigorous evaluation system in Tennessee has influenced teachers. Specifically, we ask whether teacher ratings under the new system causally affect job satisfaction. Job satisfaction among teachers has been the subject of active research for decades.

In addition to being of independent interest, job satisfaction has also been strongly linked to teacher turnover, which is a consistent area of policy focus for states and school districts across the United States. For example, Ingersoll (2001) reports that 25% of individuals who exit teaching cite job dissatisfaction as a reason for leaving, while a study focusing on novice teachers finds that "the most important influence on [teacher] retention decisions is job satisfaction" (Stockard & Lehman, 2004, p. 742).[3]

Given the rapidly changing policy landscape surrounding teacher evaluations in K–12 public schools (Steinberg & Donaldson, 2016) and the shift toward more differentiated ratings nationally (albeit slowly, see Kraft & Gilmour, 2016), a question we aim to answer is whether the new ratings in Tennessee have led to differences in job satisfaction between more and less effective teachers. This line of inquiry is consistent with the increased emphasis on quality considerations in evaluating the efficacy of personnel policies in public education (Goldhaber, Gross, & Player, 2011; The New Teacher Project, 2012; Steele, Pepper, Springer, & Lockwood, 2015). To perform our analysis, we merge ratings data for Tennessee teachers with data from a statewide survey conducted after teachers received their first-year ratings from the system. The survey includes a number of questions about professional well-being. We identify the causal effects of teacher ratings on job satisfaction by leveraging discontinuities in the way that the evaluation system converts underlying performance measures into teachers' overall ratings, as in Dee and Wyckoff (2015) and Springer et al. (2016). In particular, we estimate regression-discontinuity (RD) models that compare job satisfaction between teachers who receive similar underlying performance scores but whose scores place them on different sides of the threshold values that are used to convert underlying scores into final ratings.

Our findings indicate that teachers who receive higher ratings based on the Tennessee criteria are significantly more satisfied with work relative to otherwise similar teachers who receive lower ratings near rating thresholds. On average across the rating thresholds we study (which we describe in detail in the following), we find that a higher (lower) rating corresponds to a 0.08 to 0.09 standard deviation increase (decrease) in job satisfaction as measured by a summary satisfaction index. A heterogeneity analysis reveals that teachers' responses to their ratings are stronger at thresholds in the system that divide more extreme performance labels. This result is in line with recent evidence from Papay, Murnane, and Willett (in press), who document similar labeling-effect heterogeneity in their investigation of how labels of test score performance affect students' subsequent human capital investments. While it is intuitive that the effect of receiving a higher rating on job satisfaction for teachers is positive, to the best of our knowledge, our study is the first to confirm the causal relationship and document the effect heterogeneity. Understanding the relationship between job satisfaction and performance ratings for teachers is important in the current

education policy context as state and local education agencies increasingly consider the use of more differentiated ratings to evaluate public school teachers.

Background: Teacher Evaluations in Tennessee

As part of the federal Race to the Top (RTTT) grant competition, the United States Department of Education (USDOE) called for states and their participating school districts to improve teacher and organizational effectiveness by developing comprehensive educator evaluation systems.[4] The USDOE (2009) particularly called for systems to (a) measure student growth for each individual student; (b) design and implement evaluation systems that include multiple rating categories that take data on student growth into account as a significant factor; (c) evaluate teachers and principals annually and provide feedback, including student growth data; and (d) use these evaluations to inform decisions regarding professional development, compensation, promotion, retention, tenure, and certification. State-awarded RTTT grants have intensified efforts around the construction and implementation of performance-based measures for teacher evaluation systems, though the degree of implementation has varied considerably (U.S. General Accountability Office, 2013).

In July 2011, the Tennessee State Board of Education approved four teacher evaluation models. The models are the Tennessee Educator Acceleration Model (TEAM), Project Coach (COACH), Teacher Effectiveness Measure (TEM), and Teacher Instructional Growth for Effectiveness and Results (TIGER). All of the models follow the requirements adopted by the State Board of Education and have the same goals—to monitor teacher performance and encourage teacher development.

We omit data from the single district in Tennessee that implemented the TEM model from our study for technical reasons that we discuss in the following. Among the remaining three models—COACH, TEAM, and TIGER—the basic components of teacher evaluations are the same, although there are some differences in how the evaluations are implemented. For example, both the TEAM and TIGER models rely on announced and unannounced classroom observations while the COACH model uses unannounced observations only. There is also variability in the number and frequency of observations; TEAM requires six observations totaling more than 90 minutes for apprentice teachers and four observations totaling more than 60 minutes for tenured teachers, TIGER requires a minimum of six observations totaling more than 60 minutes, and COACH requires that all teachers be observed at least 10 times a year for a minimum of 5 minutes per observation. All three models have slightly different approaches to feedback, but all require a post-observation conference and written feedback to the teacher provided by the observer.

In terms of how performance is evaluated, state law specifies that 50% of the evaluation must be based on student achievement data in all evaluation models. Among teachers for whom individual growth measures based on how their students perform on standardized tests can be constructed, 35% of the final rating is based on student performance on these tests. For teachers outside of grades and subjects for which standardized tests are available, school- and/or district-wide growth is substituted for the individual growth measure.[5] The remaining 15% is based on additional measures of student achievement chosen through mutual agreement by the educator and evaluator. The non–achievement based components that contribute to teachers' ratings include teacher observations, student surveys, personal conferences, and a review of prior evaluations and work.

The effectiveness scores based on the aforementioned criteria range from 0 to 500 in all four evaluation models and are used to assign teachers to discrete performance categories. Denoting X as the teacher score, for all models, teachers with $X < 200$ are categorized as "significantly below expectation" (Level 1), teachers with $200 \leq X < 275$ as "below expectation" (Level 2), teachers with $275 \leq X < 350$ as "at expectation" (Level 3), teachers with $350 \leq X < 425$ as "above expectation" (Level 4), and teachers with $X \geq 425$ as "significantly above expectation" (Level 5).[6] Level 1 ratings are very rare, but ratings at the other levels are common (see Table 1). Rating reports provided to teachers include the discrete rating but not the underlying score on the 0 to 500 scale. This is useful for interpreting our findings because it means that teachers with very similar underlying scores but different discrete ratings were not provided with information to determine their closeness to the threshold. Principals receive access to evaluation ratings at the same time as teachers gain access and also do not see teachers' underlying scores.

While Tennessee law indicates that teachers' evaluation ratings will be incorporated into compensation, promotion, retention, tenure, and certification decisions, only some of these policies had been drafted and implemented at the time of the survey (spring 2013; see following), and they did not apply universally. As one example of a policy that was in place, in 2011, the Tennessee General Assembly voted to explicitly tie evaluation ratings to new tenure decisions (Public Chapter 70) by requiring teachers to receive ratings of four or five during the last two years of the pre-tenure probationary period to earn tenure. Teachers who do not receive tenure at the end of their five-year probationary period may either be rehired under a year-to-year contract or dismissed. Another example is that teachers working in disadvantaged schools under the Tennessee Achievement District program who earn a rating of three or higher can earn salary increases and/or promotions not available to teachers with lower ratings.[7] However, most teachers in Tennessee were not covered by any state policy linking ratings to rewards and sanctions at the time of our study. We return to this point when we discuss our findings in the concluding section.

Table 1
Descriptive Statistics

	All Tennessee Teachers	All Tennessee Teachers, Excluding Memphis	Survey Respondents Plus Other Initial Restrictions	Responded to at Least One Job Satisfaction Question	RD Sample, Main Bandwidth	RD Sample, Main Bandwidth, Exclude Partial Responders
Female teacher	0.80	0.80	0.82	0.82	0.81	0.81
Black teacher	0.12	0.07	0.06	0.05	0.05	0.05
White teacher	0.87	0.93	0.94	0.94	0.94	0.95
Other race	0.01	0.00	0.00	0.00	0.00	0.00
Bachelor degree	0.41	0.41	0.38	0.39	0.41	0.40
Education specialist	0.07	0.07	0.08	0.08	0.08	0.08
Master degree	0.42	0.43	0.44	0.44	0.43	0.44
Other education	0.10	0.09	0.10	0.09	0.08	0.09
Teacher experience	12.82	12.76	13.41	13.22	12.90	12.81
Group 1 (with individual growth score)	0.32	0.33	0.36	0.39	0.38	0.38
Professional license	0.85	0.86	0.89	0.89	0.88	0.88
Apprentice license	0.14	0.13	0.11	0.11	0.12	0.12
Partial-year license exemption	0.01	0.01	0.00	0.00	0.00	0.00
COACH evaluation model	0.06	0.06	0.05	0.05	0.04	0.04
TEAM evaluation model	0.82	0.91	0.91	0.92	0.92	0.92
TEM evaluation model	0.10	0.00	0.00	0.00	0.00	0.00
TIGER evaluation model	0.02	0.03	0.03	0.03	0.03	0.03
Level 1 (significantly below expectations)	0.00	0.00	0.00	0.00	0.00	0.00
Level 2 (below expectations)	0.07	0.06	0.06	0.06	0.06	0.06
Level 3 (at expectations)	0.19	0.19	0.21	0.21	0.24	0.24
Level 4 (above expectations)	0.31	0.31	0.31	0.31	0.36	0.36
Level 5 (significantly above expectations)	0.40	0.41	0.42	0.41	0.34	0.34
Percentage of minority students in school	27.47	23.04	22.54	21.96	21.98	22.04
Percentage of female students in school	47.09	48.23	48.29	48.29	48.26	48.27
Percentage of FRL students in school	53.28	54.23	54.64	54.51	55.12	55.05
N	69,796	62,867	20,330	15,437	13,266	12,169

Note. Among all Tennessee teachers, just 0.4% of teachers received an evaluation rating of 1, which indicates significantly below expectation. Group 1 teachers are those with an individual growth score as a component of the final performance measure. An education specialist degree is an advanced terminal degree for individuals who wish to go beyond the MA level but do not wish to pursue a doctorate. As noted in the text, some teachers in the "full universe sample" are missing some of the information reported in this table. Teachers with missing information are omitted from the calculations on an item-by-item basis. The data restrictions that determine these samples are listed in detail in Appendix Table A.1. COACH = Project Coach; TEAM = Tennessee Educator Acceleration Model; TEM = Teacher Effectiveness Measure; TIGER = Teacher Instructional Growth for Effectiveness and Results; FRL = free/reduced price lunch.

Empirical Strategy

We use a RD design to identify the causal impacts of ratings on teacher job satisfaction. The RD design compares teachers whose underlying performance scores are similar but who receive different ratings because of the discrete function that translates the underlying scores into final ratings. The key assumption for causal inference within the RD framework is that teachers with similar underlying scores are similar in other respects, and thus conditional on the underlying score, the discontinuous rating assignments can be viewed as effectively random (Hahn, Todd, & Van der Klaauw, 2001; Imbens & Lemieux, 2008).

The outcomes in our primary RD models are job satisfaction indices that we construct using factor analysis based on teachers' responses to 10 questions about job satisfaction on the Tennessee Race to the Top Survey. We describe the job satisfaction questions and index construction in the following. We use the indices as dependent variables in linear RD models following a "sharp" discontinuity design. Our primary specification is a "stacked" model where we estimate the average treatment effect of the higher rating across the performance thresholds 2/3, 3/4, and 4/5 (we do not study the 1/2 threshold because very few teachers in Tennessee receive a Level 1 rating as indicated by Table 1). We also consider models that separately examine each threshold to investigate the potential for effect heterogeneity at different thresholds. The models take the following form:

$$Y_i = \beta_0 + \mathbf{X_i}\beta_1 + [fS_i]\beta_2 + [fS_i * IS_i \geq T]\beta_3 + [IS_i \geq T]\beta_4 + \varepsilon_i. \quad (1)$$

In Equation 1, Y_i is the job satisfaction index; $\mathbf{X_i}$ is a vector of observable teacher and school characteristics; $f(S_i)$ is a function of the underlying score, or running variable; $I(S_i \geq T)$ is an indicator function equal to 1 if the score is above the threshold (i.e., the regression-discontinuity indicator); and ϵ_i is the error term, which we cluster at the school level.[8] The X-vector includes teacher gender, race, degree level, certification level, and experience, plus the model used to evaluate the teacher and whether the teacher had an individual growth score. It also includes controls for the shares of students at the teacher's school who are a disadvantaged minority, female, and eligible for free or reduced price lunch. Our preferred specification includes the X-vector to increase precision, but we also confirm that are findings are robust to omitting the X-vector (see Appendix D). Although we considered several functions for $f(S_i)$, in our primary models, we specify $f(S_i)$ as a simple linear function of the running variable on both sides of the discontinuity.[9] The parameter of interest is β_4, which under the RD assumptions is the causal effect of receiving a higher rating relative to a lower rating on the job satisfaction index.

An issue that arises with the stacked model is that individual teachers can in principle be identified as both treatments and controls at different thresholds. For example, a teacher with an overall score of 4 is on the

high side of the 3/4 cutoff and the low side of the 4/5 cutoff. We ensure that individual teachers are not double-counted in our main models by specifying the bandwidth around each RD cutoff at 37. Per the documentation of the score ranges in the previous section, 37 is the highest integer-value bandwidth that we can assign across all three rating thresholds and still ensure that no individual is double-counted. We also examine the robustness of our findings to narrower bandwidths in a sensitivity analysis.[10]

Data

Ratings, Administrative and Survey Data

We merge teacher ratings data with administrative data from the Tennessee Department of Education and data from an annual survey given to all Tennessee teachers as part of the FTTT initiative. The ratings data are based on teacher performance during the 2011–2012 school year. The administrative data include information on each teacher's gender, race, education, certification level, and years of experience, also from the 2011–2012 school year. The survey data are from the annual Tennessee Consortium Survey administered during the spring of 2013, after teachers received their first set of ratings from the evaluation system (ratings from the 2011–2012 school year were accessible to all teachers by February 2013, and the survey was administered in April and May of the same year). The survey is designed to improve the state's understanding of how the performance evaluation is implemented and how feedback is provided to and processed by teachers.

Table 1 provides descriptive statistics for all teachers in the Tennessee ratings database, which includes teachers regardless of whether they received a rating, side-by-side with descriptive statistics for several subsamples of teachers. The second to last column (Column 5) shows our primary analytic sample, while the other columns show various cuts of the data for illustrative purposes. In total, our primary RD sample of 13,266 teachers in Column 5, while large, represents less than a fifth of the teachers in Tennessee. The two most substantial sources of data loss revealed in Appendix Table A.1 are (a) many teachers did not fill out a survey and/or did not answer at least one job satisfaction question and (b) we cannot evaluate TEM teachers. As noted previously, we exclude TEM teachers due to TEM's unique approach to calculating overall scores, which produces a lumpy distribution of teacher performance measures that is not compatible with our regression-discontinuity research design. The TEM model was adopted only in Memphis during the 2011–2012 school year, which although just 1 of 140 Tennessee school districts in our data, accounts for approximately 10% of teachers (Memphis accounts for roughly 10% of the population in Tennessee as well). Column 2 of Table 1 provides descriptive statistics for the sample after excluding Memphis.

With regard to survey behavior, almost two-thirds of teachers did not submit a survey (see Appendix Table A.1). Column 3 of Table 1 shows the sample of teachers restricted to survey respondents (and conditional on a few other data restrictions as indicated in Appendix Table A.1). Although the response rate to the survey is not particularly low for a non-mandatory instrument (e.g., see Rockoff, Staiger, Kane, & Taylor, 2012; Stutz, 2014; Watts & Becker, 2008), the large fraction of teachers who are dropped from the sample because they did not submit a survey raises two potential concerns. First, the omission of these teachers could affect our ability to identify the rating effects using our RD design if the rating treatments also affect survey submission and completion behaviors (a related problem arises in McCrary & Royer, 2011). One way that survey submissions could be affected is if the rating discontinuities affect attrition from teaching. However, although our following tests allow for this possibility, in principle it should not be problematic for our sample because Year 1 system ratings were not finalized until the next school year had already started, as described previously. The second concern stemming from our restricted sample is that even if ratings do not affect survey submission and completion behaviors, the generalizability of our findings will be limited to the extent that teachers who submitted a survey are systematically different from other teachers in Tennessee. We address these issues in more detail in the following.

Conditional on submitting a survey, 24% of teachers either voluntarily skipped or were directed to skip the 10 job satisfaction questions.[11] We restrict our analysis to individuals who answered at least one job satisfaction question because our outcomes of interest are constructed based on teachers' answers to these questions. We have verified that excluding individuals who skipped all 10 job satisfaction questions from our analysis has no bearing on our findings qualitatively (results shown in Appendix D; also see Table 2). Column 4 shows descriptive statistics for the sample of teachers who answered at least one job satisfaction question.

The samples in Columns 4 and 5 are very similar, with the only difference being that the sample in Column 5 includes the restriction from our RD models to use a bandwidth of 37 points around the discontinuity thresholds at the 2/3, 3/4, or 4/5 cutoffs, as described in the preceding section. The most significant consequence of the bandwidth restriction is that we drop data from teachers with scores in the bottom half of the range for a rating level of 2 and the top half of the range for a rating level of 5 because they are outside of the bandwidth. As a result, our final analytic sample has fewer Level 2 and Level 5 teachers than the population of Tennessee teachers.[12]

Finally, in Column 6, we show an alternative analytic sample consisting only of individuals who answered all 10 job satisfaction questions. Our preferred approach is to include all teachers who answered at least one job satisfaction question (from Column 5) because even partial responders provide

Table 2

Regression-Discontinuity Estimates of the Effect of a Higher Rating on Survey Response Behaviors, Overall and for Each Threshold Separately

Dependent Variable	Stacked Model	Threshold 2/3	Threshold 3/4	Threshold 4/5
Any survey submission	0.007 (0.009)	0.006 (0.022)	–0.018 (0.017)	0.016 (0.011)
Survey submission and answered at least one job satisfaction question	0.004 (0.008)	0.004 (0.021)	–0.019 (0.015)	0.012 (0.010)
N	48,529	8,044	12,838	27,647

Note. Models are specified as linear probability models. Standard errors clustered at school level are reported in parentheses. The sample used for these regressions includes all individuals with complete data as shown in Row 5 of Appendix Table A.1, with the additional restriction that they are within 37 points of the discontinuity cutoffs at rating levels 2/3, 3/4, and 4/5. Summary statistics for the sample used in these regressions are shown in Appendix Table A.2.

useful information that can be exploited for our study. As described in the following, we impute missing survey answers based on observed survey answers for partial responders. However, to illustrate that the imputation is not disproportionately driving our findings, we also present results in the following using a sample that excludes partial responders. Column 6 of Table 1 provides descriptive statistics for this sample.

We now return to the concerns raised previously about the low survey response rate. First, we test whether teachers' survey submission and completion behaviors are influenced by the discontinuities that convert underlying scores to ratings. We estimate RD models for all Tennessee teachers where the dependent variable is either (a) a binary indicator for whether the teacher submitted a survey or (b) a binary indicator for whether a teacher submitted a survey *and* answered at least one job satisfaction question. The RD models we use for these tests are analogous to the model described in the preceding section and impose the 37-point bandwidth used for our main analysis.[13] Their purpose is to determine whether treatment, in this case a higher rating, impacts survey submission and completion outcomes. If so, this would suggest a source of sample selection bias in the estimates of rating effects on job satisfaction obtained from our restricted sample of teachers who submitted a survey and answered at least one job satisfaction question.[14]

Table 2 displays the estimated effects of treatment on the two aforementioned outcomes, both overall by "stacking" the discontinuities and for each discontinuity separately. The estimates are substantively small and statistically insignificant for both outcomes, overall and at each threshold

separately. We conclude that survey participation is not caused by the rating treatments, at least subject to the local interpretation of the RD estimates, which is the most relevant interpretation for informing the credibility of our main analysis. Thus, our RD estimates should not be biased by sample selection.[15]

On the issue of the generalizability of our sample, which is relevant for gauging the external validity of our findings, we return to the comparisons in Table 1. Columns 1 and 2 show that the loss of TEM teachers (Memphis) from the analytic sample influences the shares of teachers and students by race because the student population and teaching workforce in Memphis is disproportionately Black. Also, as noted previously, the RD sample includes smaller shares of Level 2 and Level 5 teachers because the bandwidth restriction cuts off the edges of the score ranges for those two scores. Outside of these differences, however, the other differences between our sample and the larger universe of Tennessee teachers are small. For example, while the teachers in our primary analytic sample (Column 5) are more experienced on average, more likely to have a master's degree, and more likely to have a professional license and these differences are statistically distinguishable given our large sample size, they are not substantive. On the whole, we interpret the small descriptive differences between the full sample of Tennessee teachers and our primary analytic sample in Table 1 as suggestive that our findings will generalize, at least to some degree, to the broader teaching population.

Turning briefly to the descriptive content of Table 1, in addition to generally describing the teaching workforce and our sample, the table also documents the prevalence of each evaluation model in Tennessee and the distribution of teacher ratings (which we also show graphically in Appendix Figure A.1). Note that the TEAM model is the predominant evaluation model in the state, and as such, our results are driven primarily by evaluation under TEAM. Regarding the overall ratings distribution, the table shows that very few Tennessee teachers receive a score that puts them at a Level 1 rating (specifically, 0.4% of all Tennessee teachers receive a Level 1 rating in Column 1, which is rounded to zero in the table). Because of the very small number of teachers who receive Level 1 ratings, we do not evaluate the effects of ratings around the 1/2 threshold. However, moving beyond Level 1, there are significant fractions of teachers who receive ratings at Levels 2 through 5, which facilitates our investigation of rating effects at the thresholds 2/3, 3/4, and 4/5. The distribution of ratings may seem somewhat generous in the absence of context. However, when compared to findings from Weisberg et al. (2009), there are clear differences. For example, in an analogous five-level rating system in Akron Public Schools, Weisberg et al. find that 60% of tenured teachers receive the highest rating and less than 1% receive one of the bottom two ratings; the analogous numbers in Table 1 are 40% and 8%, respectively (Table 1 reports on ratings

for all Tennessee teachers including untenured teachers; in results omitted for brevity, we confirm that the rating distributions are similar between tenured and untenured teachers in Tennessee, although ratings are slightly higher for tenured teachers). Similarly, in the four-level rating system in Chicago, Weisberg et al. report that almost 70% of tenured teachers receive the top rating while less than half of 1% receive a rating below satisfactory. These comparisons suggest that the rigor of the Tennessee system is substantially higher than the "industry standard" in public education.

Measuring Job Satisfaction

There are 10 questions on the survey instrument that elicit feedback from teachers regarding their satisfaction with work. In Table 3, we split the 10 questions into two groups—one group consists of questions that reference general job satisfaction and the other job satisfaction specific to the school. Answers to the questions are ordered responses, with a range of response options across questions of 3 to 7 as shown in Appendix B.

For presentational convenience, in this section we divide teacher responses to each question into just two categories: positive and negative responses.[16] To code each answer as either a positive or negative response, we collapse teachers' more detailed ordered responses as shown in Appendix B. A positive response is always coded to indicate more satisfaction with work. Table 3 shows the share of positive responses to each question across teachers by their final rating in the evaluation system. With the exception of Question 4, teachers who are assigned a higher (lower) rating indicate higher (lower) satisfaction with work, and even for Question 4, teachers with higher (lower) ratings do not indicate that they are meaningfully (more) less satisfied with work. Some of the differences across teacher ratings are modest (e.g., Question 3), and others are large (e.g., Question 7).

While the descriptive statistics in Table 3 indicate an association between teacher ratings and job satisfaction, attributing the direction of causality is not straightforward. We cannot tell if higher ratings cause greater job satisfaction or if greater job satisfaction causes higher ratings (i.e., perhaps more satisfied teachers are also more effective teachers; see Lawler & Porter, 1967; Schwab & Cummings, 1970). We overcome this identification problem with the RD models described previously, which identify the rating effect by comparing otherwise similar teachers who differ only in the final rating received, thus isolating the causal pathway from the rating to job satisfaction.

Table 4 presents a correlation matrix that documents dependence in the job satisfaction data, again based on the "positive/negative" response coding. Unsurprisingly, teacher responses are positively correlated across the job satisfaction questions. The correlations range from 0.11 to 0.61 and are typically larger within the "general" and "school-specific" sets of questions than across. The key information conveyed by the correlation matrix is

Table 3

The Proportion of Positive Responses to Each Survey Question

	Below Expectation (2)	At Expectation (3)	Above Expectation (4)	Significantly Above Expectation (5)
General satisfaction with teaching				
Question 1: Teacher prefers a higher paying job to teaching	.55	.59	.60	.63
Question 2: Teacher is less enthusiastic about teaching	.41	.44	.44	.47
Question 3: Teacher regrets teaching career choice	.46	.47	.47	.50
Question 4: Teacher plans to quit teaching	.40	.40	.39	.39
Satisfaction at school				
Question 5: Teaching in this school is not worth the stress	.65	.72	.73	.78
Question 6: Teacher believes that the staff is satisfied at this school	.60	.67	.69	.74
Question 7: Teacher likes how the things are run at this school	.57	.65	.65	.72
Question 8: Teacher would like to transfer to another school	.61	.69	.69	.74
Question 9: Teacher is tired of teaching at this school	.66	.67	.67	.70
Question 10: Teacher plans to return to this school next year	.82	.85	.86	.88

Note. Positive responses are coded as the ones that convey satisfaction with teaching. For example, for Question 6, the most positive response is *strongly agree*; for Question 1, it is *strongly disagree*. See Appendix B for more information about how positive and negative responses are coded for each question.

Table 4
Correlation Matrix for Positive Responses Across Questions

Question	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Q1	1.00									
Q2	0.42	1.00								
Q3	0.48	0.39	1.00							
Q4	0.34	0.35	0.40	1.00						
Q5	0.37	0.34	0.25	0.20	1.00					
Q6	0.23	0.26	0.19	0.16	0.48	1.00				
Q7	0.22	0.27	0.17	0.16	0.47	0.61	1.00			
Q8	0.27	0.27	0.18	0.14	0.44	0.41	0.46	1.00		
Q9	0.34	0.41	0.28	0.26	0.32	0.23	0.20	0.24	1.00	
Q10	0.16	0.15	0.11	0.13	0.31	0.25	0.27	0.36	0.13	1.00

Note. Question numbers correspond to those in Table 3. The upper box blocks off general questions, and the lower box blocks off school-specific questions.

that answers to the job satisfaction questions are interrelated, which supports the construction of summary indices to capture teacher responses holistically.

The Job Satisfaction Indices

We measure rating effects on summary indices of job satisfaction that we construct using factor analysis to consolidate the information across the 10 job satisfaction questions. By combining the information contained in all 10 survey questions, we improve statistical power in our study. We code teachers' answers to the job satisfaction questions in two different ways to construct two different indices. First, we code teachers' answers as binary positive/negative responses, as explained in the previous section and in Appendix B. This coding scheme consolidates the responses to each question into just two categories and thus results in some information loss but has the advantage of not falsely imposing a cardinal interpretation on the Likert scale survey responses. The second approach follows the larger literature using similar Likert data and assigns numerical values to the survey response outcomes (Fowler, 2013). For example, for a standard four-response question with answers (1) *strongly disagree*, (2) *disagree*, (3) *agree*, and (4) *strongly agree*, we assign the option *agree* a value of 3 and the option *strongly disagree* a value of 1. That is, *agree* is coded as indicating three times as much satisfaction as *strongly disagree*.

We denote these two coding schemes as positive/negative and numerical. Using both data coding schemes, we use the same approach to construct the job satisfaction index. As a first step in the process, we impute values for

all missing responses among our sample of teachers who answered at least one question about job satisfaction on the survey. We impute missing responses using two imputation methods: the expectation-maximization (EM) algorithm and the Markov Chain Monte Carlo (MCMC) technique (for an overview of these approaches, see Schafer, 1999). The results that we show in the following are based on data imputed using the EM algorithm, although using the MCMC technique yields similar findings. The imputed responses are continuous numeric values and can be interpreted as either (a) the likelihood of a positive response or (b) a numerical response value, depending on the coding scheme. The fact that the imputed values are continuous is perhaps unintuitive given that all actual responses take on integer values. However, mathematically, the imputation is not problematic for constructing the indices, and it allows us to use data from individuals who only partially answered the job satisfaction questions. We also show results in the following that confirm our findings are substantively similar if we drop individuals who did not submit a response to all 10 job satisfaction questions, in which case, no imputation is necessary.

The next step is to use factor analysis to consolidate the information in the 10 survey questions into a smaller number of factors that reflect underlying latent constructs of job satisfaction. Our factor analysis is based on the correlation matrix for the survey questions, which we use to gain inference about which questions or sets of questions can be grouped under common constructs. In any factor analysis, there are a number of subjective decisions that researchers must make without clear decision rules. We follow common practices established in the literature in making these decisions, leaning most heavily on Fabrigar, Wegener, MacCallum, and Strahan (1999) and Henson and Roberts (2006) to guide our choices.

The most important subjective decision noted by these and other authors is the choice of the number of factors. Standard practice is to use the magnitudes of factors' eigenvalues—which provide an indication of the variance in all of the variables accounted for by a particular factor—to select factors. Appendix Figure C.1 plots the eigenvalues from our factor analysis using each coding scheme. In both cases, the first factor is by far the most important. The second factor is marginally important, with an eigenvalue close to but below one, and all other factors have eigenvalues near zero.[17] Two common decision rules in the literature for selecting the number of factors are (a) choose all factors with eigenvalues greater than one and (b) choose all factors before the last "substantial" drop in the magnitude of eigenvalues, referred to as the "scree test." In our application, it is debatable whether the last "substantial" drop occurs after Factor 1 or Factor 2, but this is clearly the full choice set.

Although both rules are imperfect (Fabrigar et al., 1999), reasonable inference leaves just two options in our case—that we have either a one-factor or two-factor model. We pursue both of these possibilities in the

following. Appendix Table C.1 shows the loading weights from the one-factor and two-factor models (after oblique rotation for the two-factor model) using both coding schemes, which we rescale to sum to unity within factors for ease of interpretation.[18]

The job satisfaction indices are weighted averages of teachers' responses to the survey questions where the weights are the loading factors shown in Appendix Table C.1. We make several observations about the index weights. First, in the single-factor model, Questions 5 through 8—and particularly Question 5—consistently receive high weights. The high weights on these questions are noteworthy because, as we show in the following, these questions are the ones with the largest and most significant coefficients when we estimate models on a question-by-question basis. Second, in the two-factor model, the loading weights largely align the questions based on how they are grouped as previously described—into general and school-specific satisfaction measures (the exception is Question 9, which the factor analysis indicates fits better with the general set of questions, as would be predicted by the correlations in Table 4). This result is obtained despite our constructing the initial groupings without consulting the data and is reassuring. Third, the nature of the loadings in the two-factor model continues to point toward the school-specific questions as the group of questions that is most informative about job satisfaction overall (again, with the exception of Question 9) as they receive the higher weights in the significantly more important factor.

Validation of the RD Design

The RD design offers a credible approach for identifying the causal effects of teacher ratings on job satisfaction subject to several assumptions. In this section, we review and test these assumptions to provide evidence on the extent to which this approach can be useful for informing our research question.

We first examine whether the discontinuities in the data are sharp or fuzzy. Figure 1 shows the probability of a teacher receiving treatment (i.e., the higher rating) as a function of his or her underlying performance measure. In the figure, we aggregate the scores for individual teachers into 5-point bins, center them on the threshold value for the higher rating, and stack the data across all three thresholds covered by our study. The discontinuities in converting the underlying performance measures into final ratings are clearly sharp, supporting the previously specified RD design.

We also perform two general tests that are commonly used to detect potential violations of the RD assumptions. The first test examines whether there are other discontinuities in the data that align with the rating discontinuities. If other variables are discontinuous at the discontinuity thresholds, it would suggest that individuals with similar running-variable values near the cutoff are not otherwise similar.[19]

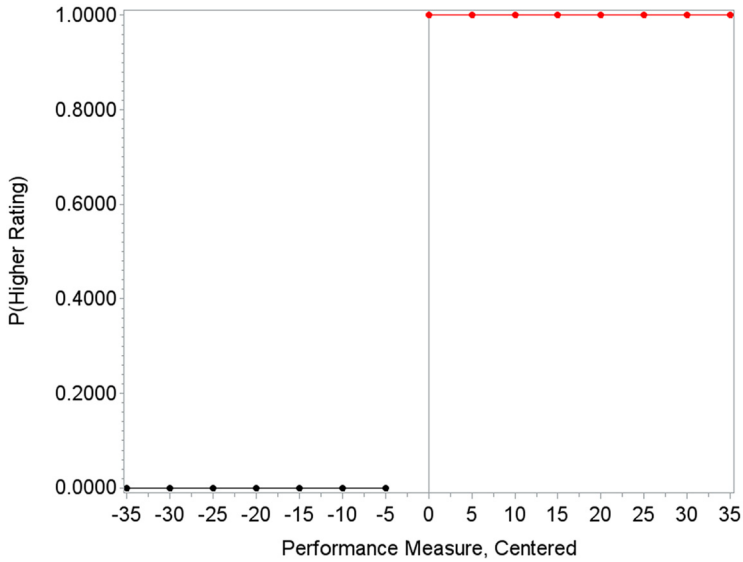

Figure 1. **Illustration of sharp discontinuities at the cut scores between levels 2/3, 3/4, and 4/5.**

Note. We aggregate teachers into 5-point bins based on their underlying performance scores. Thus, each point on the graphs denotes the probability of being assigned to the higher rating for teachers with underlying performance at that point or less than 5 points above it. As a practical matter, final ratings are determined based on underlying scores after rounding, and thus the graphs in this figure portray final ratings as a function of rounded underlying scores (see text and Appendix D for further discussion).

To determine whether other discontinuities in the data are present and align with the main discontinuities in teachers' evaluation scores, we estimate a stacked model of the following form:

$$X_i = \alpha_0 + [f(S_i)]\alpha_1 + [f(S_i) * I(S_i \geq T)]\alpha_2 + [I(S_i \geq T)]\alpha_3 + u_i. \quad (2)$$

In Equation 2, X_i is a teacher or school characteristic from Equation 1, now used as a dependent variable; all other variables and functions are specified as in Equation 1; and u_i is the error term.

Table 5 presents results from the series of RD regressions linking the rating discontinuities to teacher and school characteristics. Estimates using our primary 37-point bandwidth are shown in Column 1, and for completeness, we show estimates for a range of other bandwidths down to 5 points. Each cell shows results from a different regression.

Table 5

Regression-Discontinuity Estimates of the "Effects" of a Higher Ratings on Teacher Characteristics that Should Not Be Affected, for the Purpose of Validating the Research Design

	Bandwidth				
	37	30	20	10	5
Female	–0.028**	–0.026*	–0.041**	–0.032	–0.013
	(0.014)	(0.015)	(0.018)	(0.026)	(0.038)
African American	0.006	0.004	0.006	0.01	–0.006
	(0.009)	(0.01)	(0.012)	(0.016)	(0.022)
White	–0.005	–0.002	–0.007	–0.013	0.001
	(0.009)	(0.01)	(0.012)	(0.016)	(0.022)
BA degree	0.017	0.012	0.009	0.022	0.032
	(0.017)	(0.018)	(0.024)	(0.033)	(0.044)
MA degree	–0.015	–0.014	–0.015	–0.011	–0.049
	(0.017)	(0.019)	(0.023)	(0.033)	(0.044)
Educational specialist	0.010	0.012	0.017	0.002	0.013
	(0.010)	(0.011)	(0.014)	(0.018)	(0.026)
Teacher experience	–0.276	–0.212	–0.290	–0.682	–1.167
	(0.336)	(0.378)	(0.457)	(0.643)	(0.825)
Group 1	–0.018	–0.034*	–0.045**	–0.062*	–0.065
	(0.017)	(0.019)	(0.023)	(0.032)	(0.043)
Professional license	–0.016	–0.023*	–0.038**	–0.055**	–0.074***
	(0.011)	(0.012)	(0.015)	(0.021)	(0.028)
Percentage of minority students in school	–0.673	–0.544	–0.153	1.541	0.886
	(0.860)	(0.949)	(1.207)	(1.624)	(2.097)
Percentage of female students in school	–0.589	–1.055	–0.429	–0.255	–0.415
	(0.776)	(0.86)	(1.066)	(1.48)	(1.955)
Percentage of FRL student in school	0.046	–0.063	–0.129	–0.423	–0.355
	(0.129)	(0.134)	(0.175)	(0.296)	(0.335)
Overall *p* value	.42	>.999	.04	.44	.40
N	13,266	10,773	7,306	3,750	1,881

Note. Models are specified as linear probability models. Each estimate in each cell comes from a separate regression. Standard errors are clustered at the school level and reported in parentheses. The *p* value indicates the probability of obtaining the observed number of statistically significant coefficients by chance at the 5% level based on 3,000 bootstrap repetitions. See Table 1 for details about the variables listed here. FRL = free/reduced price lunch.

*$p < .10$. **$p < .05$. ***$p < .01$.

Focusing on the results using our primary bandwidth in Column 1, only one covariate is unbalanced—the gender indicator.[20] Moreover, with multiple hypothesis testing, some imbalance is likely to occur by chance. In order to determine the likelihood of observing a single unbalanced covariate by

chance, we use a randomized-inference test following Cullen, Jacob, and Levitt (2005) and Fitzpatrick, Grissmer, and Hastedt (2011). To perform the test, we start by splitting the analytic data set vertically, separately blocking off teachers' covariates (dependent variables) and underlying scores (independent variables). The critical feature of the vertical blocking is that it preserves the covariance structure between the variables in the *X*-vector, which is important because the covariance structure will influence the probability of observing any given number of statistically significant relationships with the real data. Next, we randomly sort the block of teacher scores, then reconnect it to the covariate block to assign each teacher a random score. We then estimate the model in Equation 2 for each covariate and store the number of covariates that are unbalanced at the 5% level under random assignment. We repeat this procedure 3,000 times to construct an empirical distribution of covariate imbalance.

In the second to last row of the table, we report the likelihood of observing the actual number of unbalanced covariates in our data by chance based on the empirical distributions. Using our preferred bandwidth, there is one unbalanced covariate, and our procedure indicates this quite likely by chance, with a p value of .42.[21] Although the estimates using the narrower bandwidths bounce around some, which is expected, the remainder of Table 5 does not indicate covariate imbalance in our data.[22] Thus, we conclude that our sample is balanced around the discontinuity thresholds. Consistent with this result, Appendix Table D.2 shows that our findings are qualitatively unaffected if we omit covariates from our job satisfaction models entirely.

Density tests are also commonly used to examine the validity of the RD design. These tests look for evidence of "bunching" of the running variable around the discontinuity and can be useful for detecting manipulating behavior. In instances where the running variable is not smoothly distributed around the discontinuity point, the concern is that the lack of smoothness could reflect unobserved differences between individuals near the threshold (i.e., the manipulation may be nonrandom). A textbook example is a test score discontinuity where a continuous score is converted to pass-fail but where students can retake the test (e.g., see Jepsen, Mueser, & Troske, 2010; Van Der Klaauw, 2002).

We report results from density tests in Figure 2 and Table 6 on a threshold by threshold basis. We do find evidence of bunching in the data; in particular, there are more teachers with scores just above the cutoff than we would expect from a smooth distribution of underlying scores at each threshold. However, it does not appear that the bunching of the running variable is the result of nonrandom manipulation by individual teacher evaluators. Instead, it appears to be driven by a system-wide, ex post rounding of teacher scores. For example, although 425 is the official threshold for earning a final rating of 5 in the system, teachers with scores close to but just

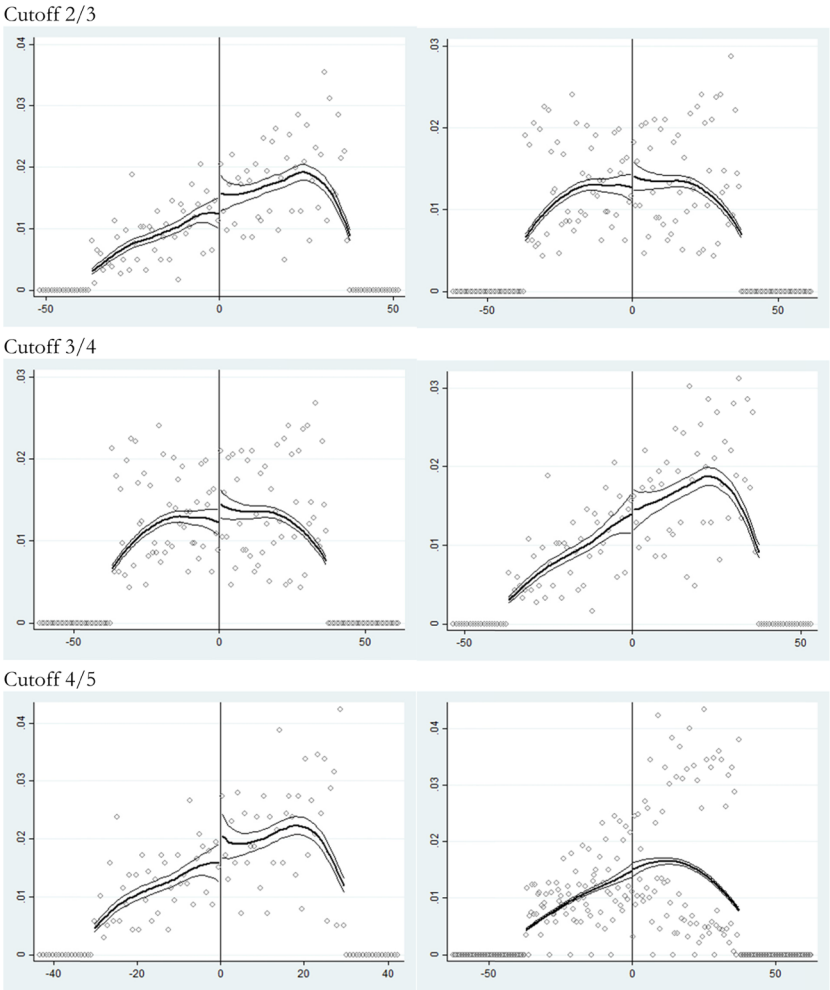

Figure 2. **Density test results at each rating threshold, (left) with and (right) without rounding.**

below this threshold were, without exception, assigned ratings of 5 as well. It is not surprising that this type of rounding would cause bunching just above the threshold.

If we perform the density tests based on hypothetical teacher ratings without allowing for the rounding that occurred in practice, the bunching becomes smaller and statistically insignificant at each threshold. This result

Table 6
Density Tests at Each Threshold Using McCrary's Method

	Cutoff 2/3	Cutoff 3/4	Cutoff 4/5
Using actual assignment variable	.254*	.178*	.180***
	(.147)	(.091)	(.054)
Using non-rounded assignment variable	.027	.112	.007
	(.135)	(.092)	(.059)

Note. Standard errors in parentheses.
*$p < .10$. ***$p < .01$.

is presented visually in Figure 2 with the side-by-side graphical illustrations of the density tests with and without accounting for rounding; it is also shown numerically in Table 6. The implication is that the failure of the density test in our data is driven by the system-wide rounding of teacher ratings and not the nonrandom manipulation of scores for individual teachers by evaluators. It is the latter scenario, which is conceptually similar to the previously described canonical test score example, that would be problematic for identification in our study. Indeed, McCrary (2008) notes that failure of the density test alone does not provide sufficient grounds to reject the validity of an RD research design. Inference depends on the source of the failure of the test, which in our case seems unlikely to be cause for concern given that the rounding occurred without case-by-case discretion. As noted, in the empirical analysis, we use teachers' ratings based on their actual, rounded scores (that is, the ratings teachers actually received), which is why all discontinuities shown in Figure 1 are sharp.[23]

Results

Graphical Evidence

Figure 3 shows our primary results graphically for the positive/negative and numerical job satisfaction indices using the single-factor model. Again, each index is a weighted average of teachers' responses to the job satisfaction questions where the weights are from the factor analysis as shown in Appendix Table C.1. A higher index value indicates more job satisfaction. Like in Figure 1, we report average index values for teachers within 5-point bins of the underlying evaluation scores and center the stacked results around the threshold value for the higher rating. Regression lines from the primary data analysis along with 95% confidence intervals are also shown. Using both index constructs, a visual jump in the level of job satisfaction is apparent at the stacked threshold for the higher rating.

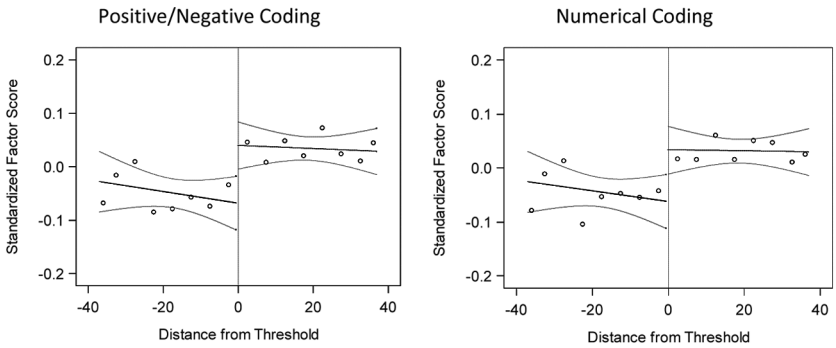

Figure 3. **Illustration of regression-discontinuity results for the single-factor job satisfaction indices using the positive/negative and numerical coding schemes.**
Note. Regression lines (solid lines) and 95% confidence intervals (dashed lines) are reported along with binned outcome data. We aggregate teachers into 5-point bins based on their underlying performance scores so that each point on the graphs denotes the value of the job satisfaction index for teachers with underlying performance at that point or less than 5 points above it.

Effects of Higher Ratings on the Job Satisfaction Indices

Table 7 presents results from our main regressions as shown in Equation 1 using each coding scheme and single-factor and two-factor indices.[24] All satisfaction indices are standardized to have a mean of zero and variance one so that the coefficients can be interpreted as showing effects in standard deviation units. Our primary results are shown in Column 1 using the full sample, including the teachers for whom some survey answers are imputed (the single-factor results in Column 1 correspond to the graphs in Figure 3). Column 2 shows results when we drop partial responders entirely rather than impute missing answers. In Columns 3 and 4, we replicate these analyses using teachers evaluated under the TEAM model only. As shown in Table 1, most teachers in our analytic sample are evaluated under TEAM, which is the predominant evaluation model in Tennessee. However, one might still worry that a correlation between model selection and outcomes could be confounding the analysis.[25] The TEAM-only regressions are presented to assuage this concern.

The results in the table confirm that the visual gaps in Figure 3 for the single-factor index are statistically significant and robust to basic data adjustments. Furthermore, there are similar statistically significant gaps using the two-factor indices. In the two-factor models, the results are always stronger for the first factor, which is notable given that the factor analysis indicates

Table 7
The Effect of a Higher Rating on Job Satisfaction

	Higher Rating	Higher Rating No Imputation	Higher Rating TEAM Only	Higher Rating TEAM Only No Imputation
Positive/negative coded index				
Single-factor model				
Index, single-factor weights	0.094***	0.083**	0.116***	0.094**
	(0.035)	(0.037)	(0.036)	(0.038)
Two-factor model				
Index, Factor 1 weights	0.098***	0.087**	0.107***	0.091**
	(0.035)	(0.037)	(0.036)	(0.038)
Index, Factor 2 weights	0.073**	0.063*	0.104***	0.081**
	(0.034)	(0.036)	(0.035)	(0.037)
Numerically coded index				
Single-factor model				
Index, single-factor weights	0.081**	0.076**	0.108***	0.092**
	(0.035)	(0.037)	(0.036)	(0.038)
Two-factor model				
Index, Factor 1 weights	0.089***	0.083**	0.102***	0.090**
	(0.034)	(0.036)	(0.035)	(0.037)
Index, Factor 2 weights	0.055	0.052	0.089**	0.073*
	(0.035)	(0.037)	(0.036)	(0.038)
N	13,266	12,169	12,245	11,229

Note. The index variables that are used as dependent variables in these regressions are weighted averages of teacher responses to the survey questions, with the weights determined by factor analysis and reported in Appendix Table C.1. Higher index values indicate more job satisfaction. Each estimate in each cell comes from a separate regression. The values in each cell are linear regression coefficients with standard errors clustered at the school level in parenthesis. The "no imputation" sample drops all teachers who did not answer all 10 survey questions rather than imputing the missing answers. The TEAM only sample drops all teachers in districts that use an evaluation model other than TEAM. In results omitted for brevity, we also estimated two-dimensional clustered models following Lee and Card (2008) due to minor lumpiness in our data, and our standard errors are nearly identical—they rise by 0.000 to 0.002 points, depending on which outcome variable we use as the dependent variable. TEAM = Tennessee Educator Acceleration Model.
*$p < .10$. **$p < .05$. ***$p < .01$.

that the first factor is significantly more informative. Focusing on the results in Column 1 from the single-factor models, our estimates indicate rating effects of roughly 0.08 to 0.09 standard deviations of job satisfaction as measured by these indices.

Table 8
Effects of Higher Ratings on Job Satisfaction, Question-by-Question Ordered-Response Models

Dependent Variable	Higher Rating
General satisfaction with teaching	
Question 1: Teacher prefers a higher paying job to teaching	1.094 (0.075)
Question 2: Teacher is less enthusiastic about teaching	1.088 (0.070)
Question 3: Teacher regrets teaching career choice	1.101 (0.067)
Question 4: Teacher plans to quit teaching	1.056 (0.067)
Satisfaction at school	
Question 5: Teaching in this school is not worth the stress	1.154** (0.078)
Question 6: Teacher believes that the staff is satisfied at this school	1.119* (0.073)
Question 7: Teacher likes how the things are run at this school	1.197*** (0.078)
Question 8: Teacher would like to transfer to another school	1.140** (0.076)
Question 9: Teacher is tired of teaching at this school	1.051 (0.068)
Question 10: Teacher plans to return to this school next year	1.162 (0.109)
N	13,266

Note. Question numbers correspond to those in Table 3. Models are specified as ordered logistic regressions. The sample size reported at the bottom of the table indicates the full sample, but individuals with missing answers to specific questions are dropped in the regression for that question (because imputed values cannot be assigned ordinal values). Standard errors clustered at the school level are reported in parentheses.
*$p < .10$. **$p < .05$. ***$p < .01$.

Effect of the Higher Rating for Each Job Satisfaction Question

We also estimate the effect of the higher rating for each question separately using ordered-response logistic regression models that are otherwise analogous to our main specification. We code the ordered responses using the same numerical values for the answers to each question as we use to construct the numerical index but note that the ordered-response model does not improperly impose a cardinal interpretation on the data.[26]

We show the results in Table 8. Each cell reports an odds ratio, with values above 1.0 indicating that a higher rating increases the likelihood of a more positive ordered response and vice versa for values below 1.0. The standard errors in parenthesis can be used to determine whether each odds ratio is significantly different from 1.0.

The question-by-question analysis is consistent with our index-based results in showing that higher ratings improve job satisfaction. Note that all of the odds ratios in Table 8 are above 1.0, and several of the coefficients are statistically significant. As noted previously, the question-by-question results are strongest for Questions 5 through 8. These questions ask whether teaching at the school is worth the stress, whether the respondent thinks the

staff at the school is satisfied, whether the teacher likes how things are run at the school, and whether the teacher would like to transfer to a new school. These questions consistently receive high weights in the job satisfaction indices as determined by the factor analysis (see Appendix Table C.1).

Robustness and Extensions

We examine the robustness of our findings in several ways. First, we examine sensitivity to alternate bandwidth specifications around the discontinuity thresholds. Our main results in Table 7 use the largest integer-value bandwidth that we can apply consistently across all three thresholds without double-counting individuals, which is 37 points per the previous discussion. In Table 9, we reduce the bandwidth around each threshold to 30, 20, 10, and then 5. Consistent with the reduction in the size of our analytic sample, when we shrink the bandwidth, our standard errors rise (at the smaller bandwidths, they rise dramatically, making the estimates much less informative), but our results are qualitatively unaffected.

Next, in Table 10 we return to using the full 37-point bandwidth and allow for effect heterogeneity by estimating separate models for each threshold. The disaggregated estimates reveal that our results are strongest at the 4/5 threshold. However, this is at least partly driven by the fact that our sample size is largest at this threshold (scores of 4 and 5 are the most common in the Tennessee system), and thus our standard errors are smallest. The point estimates at the 2/3 threshold are similar to or larger than the point estimates at the 4/5 threshold but estimated less precisely. The heterogeneity analysis uncovers no evidence to suggest that a higher rating at the 3/4 threshold affects job satisfaction.[27] Our finding that the job satisfaction effects are concentrated at thresholds that divide more extreme labels in the Tennessee system aligns with related evidence from Papay et al. (in press), who study how students respond to test score labels.

In Table 11, we report on a third robustness check examining the extent to which our findings are influenced by the fact that some teachers may not have been treated. Specifically, although all teachers were provided access to their ratings online, approximately 13% of individuals in the analytic sample indicate on the survey that they did not see their ratings.[28] We have not handled these teachers differently in the analysis thus far. However, the fact that these teachers may not have seen their ratings is likely to attenuate our results if our findings are driven primarily by psychological factors associated with receiving a higher rating, in which case, these teachers were never actually treated. Of course, if the mechanism driving our findings is unobserved career benefits of higher ratings, then these teachers would still be treated even if they did not see their own ratings—for example, they might be favored in requests from their principals even if they don't know it is because of their high ratings.

Table 9
Results With Narrower Bandwidths

	Bandwidth				
	37 (Main Results)	30	20	10	5
Positive/negative coded index					
Single-factor model					
Index, single-factor weights	0.094***	0.096**	0.091*	0.111	0.138
	(0.035)	(0.039)	(0.048)	(0.069)	(0.096)
Two-factor model					
Index, Factor 1 weights	0.098***	0.096**	0.076	0.089	0.065
	(0.035)	(0.039)	(0.048)	(0.068)	(0.095)
Index, Factor 2 weights	0.073**	0.078**	0.096**	0.127*	0.213**
	(0.034)	(0.038)	(0.046)	(0.066)	(0.091)
Numerically coded index					
Single-factor model					
Index, single-factor weights	0.081**	0.076*	0.086*	0.107	0.169*
	(0.035)	(0.039)	(0.048)	(0.069)	(0.096)
Two-factor model					
Index, Factor 1 weights	0.089***	0.083**	0.084*	0.108	0.117
	(0.034)	(0.038)	(0.047)	(0.068)	(0.094)
Index, Factor 2 weights	0.055	0.053	0.074	0.098	0.209**
	(0.035)	(0.039)	(0.048)	(0.067)	(0.093)
N	13,266	10,773	7,306	3,750	1,881

Note. The first column replicates the main results in Table 7 with a bandwidth of 37 for ease of comparison. The index variables that are used as dependent variables in these regressions are weighted averages of teacher responses to the survey questions, with the weights determined by factor analysis and reported in Appendix Table C.1. Higher index values indicate more job satisfaction. Each estimate in each cell comes from a separate regression. Standard errors clustered at the school level in parenthesis.
*$p < .10$. **$p < .05$. ***$p < .01$.

To examine this issue, in Table 11. we report results from the following regression:

$$Y_i = \delta_0 + \mathbf{X_i}\delta_1 + [f(S_i)]\delta_2 + [f(S_i) * I(S_i \geq T)]\delta_3 + [I(S_i \geq T)]\delta_4 + U_i\delta_5 + U_i * [I(S_i \geq T)]\delta_6 + e_i. \quad (3)$$

Equation 3 adds two new terms compared to what we show in Equation 1: an indicator for not observing the rating, U_i, and an interaction between U_i and whether the higher rating was received. The coefficient δ_4 indicates the effect of the higher rating for the sample that observes the rating, while (δ_4 +

Table 10
Rating Effects at Different Discontinuity Thresholds

	Level 2-3	Level 3-4	Level 4-5
Positive/negative coded index			
Single-factor model			
Index, single-factor weights	0.143*	–0.016	0.117***
	(0.084)	(0.048)	(0.039)
Two-factor model			
Index, Factor 1 weights	0.167**	–0.013	0.116***
	(0.083)	(0.048)	(0.039)
Index, Factor 2 weights	0.081	–0.008	0.098***
	(0.081)	(0.047)	(0.038)
Numerically coded index			
Single-factor model			
Index, single-factor weights	0.098	–0.006	0.100**
	(0.084)	(0.049)	(0.039)
Two-factor model			
Index, Factor 1 weights	0.135*	0.001	0.105***
	(0.082)	(0.049)	(0.04)
Index, Factor 2 weights	0.037	–0.004	0.074*
	(0.082)	(0.048)	(0.039)
N	3,957	7,992	9,309

Note. The index variables that are used as dependent variables in these regressions are weighted averages of teacher responses to the survey questions, with the weights determined by factor analysis and reported in Appendix Table C.1. Higher index values indicate more job satisfaction. Each estimate in each cell comes from a separate regression. Standard errors clustered at the school level in parenthesis.
*$p < .10$. **$p < .05$. ***$p < .01$.

δ_6) indicates the effect for the sample that does not observe the rating.[29] We show estimates of δ_4 and δ_6 in Table 11. For brevity, we report results from Equation 3 using the single-factor index models only. We obtain substantively similar results using the two-factor models as well.

The results are consistent with the hypothesis that those who report not seeing their rating are attenuating the estimates we presented earlier. In particular, our estimate of δ_4 is bigger than our corresponding estimate of β_4 (from Table 7), and the summation of ($\delta_4 + \delta_6$) indicates that individuals who did not see their ratings are essentially behaving as if they were not treated (the summed coefficients are nominally negative in each model but not statistically significant from zero in either case). Thus, we conclude that attenuation bias is reduced when we focus on individuals who report seeing their ratings.

Finally, in Table 12, we estimate our main models but restrict the sample to tenured teachers outside of schools in the highest poverty quintile in

Table 11
Accounting for Individuals Who Report Not Seeing Their Ratings

	Higher Rating
Positive/negative coded index	
Single-factor model	
Index, single-factor weights	0.116***
	(0.036)
Index interacted with unobserved rating	–0.159***
	(0.051)
Numerically coded index	
Single-factor model	
Index, single-factor weights	0.102***
	(0.035)
Index interacted with unobserved rating	–0.158***
	(0.052)
N	13,266

Note. This table replicates the results from Table 7 but adds an interaction term to the model between the higher rating indicator and an indicator for the teacher reporting not having seen the rating. As noted in the text, the model also includes a baseline indicator for teachers who did not observe their rating so that the interaction coefficient is identified. The two-factor model results are similar to what is reported here, but the results are omitted for brevity. The index variables that are used as dependent variables in these regressions are weighted averages of teacher responses to the survey questions, with the weights determined by factor analysis and reported in Appendix Table C.1. Higher index values indicate more job satisfaction. Each estimate in each cell comes from a separate regression. Standard errors clustered at the school level in parenthesis.
*$p < .10$. **$p < .05$. ***$p < .01$.

Tennessee (as measured by the student share eligible for free/reduced price lunch). Teachers in this restricted sample should not be influenced by the statewide policies associated with teacher ratings described in the Background section. The results in Table 12 are very similar to what we show in Table 7, which suggests that the statewide policies associated with teacher ratings are not driving our findings.

Discussion and Conclusion

The process of implementing new, more rigorous teacher evaluations is moving forward or being considered in a number of states and school districts across the United States. Tennessee has been at the forefront of education policy in this regard, and we use the Tennessee experience to understand how rigorous evaluations influence job satisfaction among teachers. Specifically, we ask whether the assignment of ratings itself affects job satisfaction. Using a regression-discontinuity design that locally identifies

Table 12
Effects for Tenured Teachers Outside of the Highest Poverty Schools

	Higher Rating
Positive/negative coded index	
Single-factor model	
Index, single-factor weights	0.101***
	(0.038)
Two-factor model	
Index, Factor 1 weights	0.107***
	(0.037)
Index, Factor 2 weights	0.075**
	(0.037)
Numerically coded index	
Single-factor model	
Index, single-factor weights	0.084**
	(0.037)
Two-factor model	
Index, Factor 1 weights	0.104***
	(0.036)
Index, Factor 2 weights	0.047
	(0.037)
N	11,300

Note. These estimates are comparable to the estimates in Column 1 of Table 7. The index variables that are used as dependent variables in these regressions are weighted averages of teacher responses to the survey questions, with the weights determined by factor analysis and reported in Appendix Table C.1. Higher index values indicate more job satisfaction. Each estimate in each cell comes from a separate regression. Standard errors clustered at the school level in parenthesis.
**$p < .05$. ***$p < .01$.

the effects of higher ratings near thresholds that convert teachers' underlying scores into discrete final ratings, we show that job satisfaction is causally affected by ratings. Teachers who receive higher (lower) ratings indicate that they are more (less) satisfied in teaching than otherwise similar teachers who receive lower (higher) ratings. The magnitudes of the effects that we uncover are nontrivial: On average, receiving a higher (lower) rating increases (decreases) job satisfaction for teachers in Tennessee by roughly 0.08 to 0.09 standard deviations of the distribution of job satisfaction as measured by our job satisfaction indices. Our findings are more pronounced at thresholds between more extremely labeled performance levels. In particular, they are largest at the threshold between labels of *above expectation* and *significantly above expectation* (with the latter being the highest possible rating) and at the threshold between labels of *below expectation* (Level 2) and *at expectation* (Level 3).

Some formal statewide policies were in place in Tennessee to encourage teachers to value high ratings at the time of our study. For example, non-tenured teachers were required to receive ratings of four or five during the last two years of the pre-tenure probationary period to earn tenure, and teachers working in disadvantaged schools under the Tennessee Achievement District program who earned a rating of three or higher were eligible for salary increases and/or promotions not available to teachers with lower ratings. However, these formal policies do not appear to drive our findings, leading us to hypothesize that other local informal factors and/or psychological labeling factors are responsible. As an example of a local informal factor, note that even in the absence of an explicit statewide policy, teachers may expect schools and districts to favor higher-rated teachers when making transfer and promotion decisions (Bates, 2015). Our larger findings at the more extreme labeling thresholds align with similar evidence from Papay et al. (in press), who study how test score performance labels affect students' decisions to invest in human capital. A comparison of responses by teachers who did and did not see their ratings (Table 11) is consistent with the rating labels being more important than any local career rewards and sanctions that might accumulate because of the ratings; any such rewards and sanctions should in principle accrue to teachers regardless of whether they see their own ratings.

There are a number of potential theories of action that underlie the move toward more differentiated teacher ratings in Tennessee and elsewhere. It may be a primary objective to generate differential satisfaction between more and less effective teachers, perhaps with the longer-term goal of influencing attrition and thus workforce composition in a way that improves student achievement (Adnot, Dee, Katz, & Wyckoff, 2016; Kraft, 2015; The New Teacher Project, 2012). In this case, given the strong link between job satisfaction and retention established in previous research (Ingersoll, 2001; Stockard & Lehman, 2004), our findings are consistent with progress in Tennessee on this objective (subject to the local interpretation of the estimates, which we discuss in the next paragraph). Or, our findings might simply reflect an unintended consequence (or second-order objective) of a rating system with a different primary goal, such as to improve teacher performance by providing more differentiated and productive feedback and support to teachers (e.g., Alonso & Lewis, 2001; Papay et al., 2016; Taylor & Tyler, 2012). Either way, our results merit consideration as teacher evaluation policies continue to evolve in the United States.

A caveat to our study is that while the RD design lends credence to the internal validity of the findings, our results are identified from and thus may be specific to teachers near the rating thresholds. Given that the teachers in our study did not know how close they were to the thresholds because they did not receive their underlying scores, it is perhaps reasonable to expect teachers close to and far from the thresholds to respond similarly.

However, it is also possible that the effect of a rating on job satisfaction depends on the extent to which it is a "surprise" to the teacher, and teachers farther from the thresholds may be less likely to be surprised. If this is the case, while our key parameters of interest identified using teachers near the thresholds would still be estimated accurately, the implications of our study would change. Specifically, the effects of ratings on job satisfaction throughout the workforce would be much smaller if job satisfaction outcomes for teachers far from the thresholds were less affected. We cannot test the extent to which some teachers were more surprised by their ratings than others directly with our data because we do not have information on teachers' rating expectations but note that this will be less of a concern if teachers have inaccurate and/or imprecise knowledge of their effectiveness in general. Given that there is also noise in the ratings themselves, it may be that many teachers have inaccurate and/or weak rating expectations, which would lead to "surprises" throughout the workforce (see Rothstein, 2015). One factor that may contribute to the teachers in our study having limited information about their own effectiveness is that they had just received their first ratings from the new system. Prior to its implementation, teachers were assessed under the Framework for Evaluation and Professional Growth (FEPG) in Tennessee, and the vast majority of teachers likely received very positive performance ratings with very low variance under the FEPG.[30] An interesting area for future research will be to examine how teachers' expectations about their ratings evolve over time under new, more informative systems like the one we evaluate in Tennessee.

Perhaps a more important outstanding question that cannot be resolved by our study is how the implementation of the Tennessee evaluation system has affected teacher well-being overall. We do not have pre-implementation survey data to determine the overall impact of the new evaluation system on job satisfaction. A concern is that all teachers may be less satisfied with teaching post-implementation, in which case, the differentiating effects of ratings that we identify could be swamped by the "tide" of an overall negative effect. Of course, we also cannot rule out a positive overall effect of the more rigorous evaluation system, and in fact, available evidence does not seem to suggest that job satisfaction among teachers on the whole is adversely affected by increased accountability (Grissom, Nicholson-Crotty, & Harrington, 2014). However, we hesitate to infer too much based on prior evidence because previous versions of accountability have looked different than what is emerging in new teacher evaluation systems like the one that we study in Tennessee. The question of the total workforce effect is one that merits attention in future research as additional rigorous evaluation systems come online and more data become available.

Notes

This study was supported by the Tennessee Consortium on Research, Evaluation and Development (the Consortium) at Vanderbilt University's Peabody College, which is funded by the State of Tennessee's Race to the Top grant from the United States Department of Education (grant No. S395A100032). We appreciate helpful comments and suggestions from Dale Ballou, Colleen Heflin, Peter Mueser, Michael Podgursky, and Nate Schwartz. We would also like to acknowledge the many individuals at the Consortium and Tennessee Department of Education for providing data and expert insight to conduct our analysis, in particular, Susan Burns, Sara Heyburn, Trish Kelly, Erin O'Hara, and Matthew Pepper. The usual disclaimers apply.

[1]A number of states have enacted legislation mandating performance-based evaluations for teachers, and high stakes have been attached in some cases—examples include Senate Bill 736 in Florida (2011) and House Bill 1001 in Colorado (2012). Similar legislation is being considered or in the process of being implemented in other states, including Michigan and Pennsylvania. Some large school districts are also independently developing performance-based teacher evaluations. The Houston Independent School District (Shifrer, Turley, & Heard, 2013), L.A. Unified School District (Strunk, Weinstein, & Makkonnen, 2013), Pittsburgh Public School District (Chute, 2013), and Washington, D.C., Public School District (Arcaira et al., 2013) are examples.

[2]According to the United States Government Accountability Office (GAO, 2013), Tennessee was one of only six Race to the Top (RTTT) states to fully implement both their teacher and principal evaluation systems by 2012–2013.

[3]These studies are part of a large, long-standing literature on teacher job satisfaction and turnover (e.g., also see Borg & Riding, 1991; Chapman & Lowther, 1982; Guarino, Santibanez, & Daley, 2006; Johnson, 2006; Johnson, Kraft, & Papay, 2012; Liu & Ramsey, 2008; The New Teacher Project, 2012). In addition to noting that job satisfaction is of independent interest even in the absence of a turnover, we do not directly study turnovers here because an immediate-term turnover analysis would likely understate the importance of rating effects on turnover if some turnover responses to changes in satisfaction happen with a lag. For example, it may be that changes in job satisfaction lead to changes in exit behavior but that it takes several years for a change in job satisfaction to manifest itself in the form of an exit. The turnover studies by Ingersoll (2001) and Stockard and Lehman (2004), which are based on retrospective teacher surveys, are in a better position to examine the variety of factors that build up over time to result in a turnover, and these studies identify job satisfaction as a significant factor that influences teacher exits.

[4]RTTT is a competitive grant program created under the American Recovery and Reinvestment Act of 2009. RTTT provides incentives for states to reform K–12 education in such areas as turning around low-performing schools and improving teacher and principal effectiveness. To date, the United States Department of Education (USDOE) has awarded 19 states over $4.35 billion to implement RTTT reforms (Tennessee's RTTT award was for $501 million). These states serve approximately 22 million students and employ 1.9 million teachers in 42,000 schools, representing roughly 45% of all K–12 students and 42% of all low-income students (USDOE, 2014).

[5]Approximately one-third of teachers in Tennessee have an individual growth measure (see Table 1).

[6]In practice, scores just below the consequential threshold values were rounded up so that teachers who ended up just below each threshold were awarded the higher rating without exception. For example, teachers with scores close to but just below 425 received a rating of 5 (i.e., significantly above expectation). As we illustrate in the following, the discontinuities that convert the underlying scores into final ratings are sharp after accounting for the rounding (see Figure 1). We use the actual threshold values (i.e., allowing for the rounding) in the empirical work that follows. More discussion of the rounding of scores is provided in the Validation of the RD Design section.

[7]See http://achievementschooldistrict.org/wp-content/uploads/2012/12/ Achievement-Teacher-Compensation-Summary-Feb-2013.pdf.

[8]The running variable is not perfectly continuous due to some discreteness in teachers' scores on the subcomponents. The end result is that the values of the running variable cluster around 0.5-unit intervals throughout the range of possible scores. Although the discreteness in the running variable is not egregious in our application by any means, in results omitted for brevity, we investigate its implications by estimating variants of our models where we used two-dimensional clustering at the school level and the 0.5–unit interval level, as suggested by Lee and Card (2008). The standard errors from the alternatively clustered models are very similar to what we report in the following and do not alter inference from our analysis in any way.

[9]We use the Akaike Information Criterion (AIC) test to determine the polynomial order for our primary specification. Adding higher order polynomial terms (up to quartic) of the running variable to the models does not influence our results qualitatively.

[10]In results that we do not report for brevity, we also verify that our findings are qualitatively robust to re-weighting the observations so that teachers with scores closer to the discontinuity thresholds receive higher weights than those farther away within the bandwidth.

[11]Among this group, about half skipped the questions as directed by the survey due to a position change, and the other half voluntarily skipped the questions.

[12]The data loss for Level 2 and Level 5 teachers due to the 37-point bandwidth can be readily seen by referencing the score ranges for each rating level reported in the Background section. By dropping some Level 2 and some Level 5 teachers, the bandwidth restriction improves comparability among teachers in the analytic sample with some potential consequences for generalizability.

[13]The sample used for this regression includes all Tennessee teachers who were not observer teachers, not Teacher Effectiveness Measure (TEM) teachers, had full demographic data, and had a final rating assigned. These data categories are listed in Appendix Table A.1. Also in that appendix, we provide descriptive statistics for this sample that are analogous to what we shown in Table 1. In a sensitivity analysis omitted for brevity, we also verify that our findings in Table 2 are not sensitive to bandwidth adjustments.

[14]Our approach follows that of McCrary and Royer (2011), who encounter a related problem in their investigation of the effects of female education on fertility and infant health.

[15]As noted previously, one reason that teachers did not submit a survey is that they left the Tennessee teaching workforce. These teachers are incorporated into the results in Table 2 in the sense that they are treated no differently than teachers who remained in the workforce but did not submit a survey. For the purpose of gaining inference about identification, it is not necessary to differentiate between the various reasons that teachers did not submit a survey. Also recall that the ratings for which we identify effects on job satisfaction were not made available to teachers until well after the subsequent school year had started.

[16]We provide detailed information on the share of responses to each question for each ordered option, overall and by teacher rating level, in Appendix B.

[17]The eigenvalue plot shows that the reduced correlation matrix is not positive definite (i.e., some eigenvalues are negative). This may seem odd given the intuitive (albeit approximate) interpretation of the eigenvalues as measures of variance. However, small negative eigenvalues like the ones shown in Appendix Figures C.1 and C.2 are not uncommon in factor analysis (Lorenzo-Seva, 2013). The reason that negative eigenvalues are typically obtained is that the factor analysis is analyzing only the common variance, which is less than the total variance.

[18]See Browne (2010) and Fabrigar, Wegener, MacCallum, and Strahan (1999) for discussions of the merits of oblique rotation over orthogonal rotation.

[19]Although researchers can overcome the direct threat by controlling for violating covariates in a regression, if discontinuities in observables emerge, then it raises the concern that there are other unobserved discontinuities as well.

[20]The coefficients reported in Table 5 are generally small in magnitude. To give a brief sense of this, and focusing on the results in Column 1 using our preferred bandwidth, the coefficients on the indicators for being female, African American, and having an MA

degree correspond to "effect" sizes of roughly 0.07, 0.03, and 0.03 standard deviations of these variables, respectively. The minority, gender, and free/reduced price lunch school share coefficients correspond to "effect" sizes of roughly 0.03, 0.01, and < 0.01 (0.0008) standard deviations of these variables.

[21]In results omitted for brevity, we also perform the balancing test for each discontinuity threshold separately and obtain qualitatively similar results.

[22]We do obtain one p value less than .05 using the 20-point bandwidth in Table 5, but this result is isolated, and there is no indication elsewhere in Table 5 or in Appendix Table D.2 that covariate imbalance is problematic for our analysis. In interpreting the small p value at the 20-point bandwidth, it is also important to recognize that as we expand the number of tests at different bandwidths, we are more likely to obtain some small p values by chance.

[23]Even with the rounding issue, it is correct to interpret the discontinuities in our data as "sharp"—in contrast to what we observe with the Tennessee data, a fuzzy discontinuity occurs when individuals with the same running-variable values can be assigned to different discrete performance categories with a probability above zero.

[24]We suppress the coefficients on the control variables, running variable, and running-variable interactions for brevity. These coefficients for our primary specification using each coding scheme and single-factor indices are reported in Appendix Table D.3.

[25]One may also be concerned whether there are observed differences in the scores of models where classroom observations are announced and those where classroom observations are unannounced. McClellan and Donoghue (2013) addressed this issue in detail, concluding that there is little meaningful difference between the announced and unannounced observations in Tennessee.

[26]As an example, consider Question 4 in Appendix B. Although there are seven different possible answers to this question, in both the numerical index construction and the ordered-response models, we consolidate these seven answers into just four answers. The reason is that it is not straightforward to assign an order between several of the options, and thus we group them together (but we assign indistinguishable answers as a group in an appropriate place in the order or responses).

[27]We have also estimated threshold-specific models with wider and narrower bandwidths as well. Our results are qualitatively unaffected.

[28]Instructions to teachers for how to obtain their ratings can be found at http://teamtn.org/wp-content/uploads/2013/08/Teacher-Access-Guide-for-CODE_July_2014.pdf.

[29]Equation 3 could be further modified to allow for a differential running-variable effect by whether the rating is observed, but our preferred specification identifies the relationships between the running variable and job satisfaction without distinguishing by type. As a practical matter, the results are very similar if we fully interact the model.

[30]See Tennessee State Board of Education (2004) for more information on the Framework for Evaluation and Professional Growth.

[31]The survey responses are truncated in the raw data in a way that make it impossible to distinguish between answers (b) and (c), and thus we do not report answer shares for these two options. Together, they account for 27.06% of responses. This coding issue has no bearing on our findings because answers (b) and (c) are coded the same in both the positive/negative and numerical coding schemes.

References

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2016). *Teacher turnover, teacher quality, and student achievement in DCPS* (NBER Working Paper No. 21922). Chicago, IL: NBER.

Alonso, P., & Lewis, G. B. (2001). Public service motivation and job performance. *The American Review of Public Administration*, *31*(4), 363–380.

Arcaira, E., Birman, B., Coleman, S., Dunlop, E., Feuer, M., Freund, M., Glazerman, S., Hannaway, J., Harding, H., MacFarlane, J., Nesin, T., Ozek, U., Palmiter, A.,

Turnbull, B., & Walsh, E. (2013). *Evaluation of the DC Public Education Reform Amendment Act (PERAA): Report No. 2, School Year 2011–2012*. Washington, DC: George Washington University.

Bates, M. (2015). *Public and private learning in the market for teachers: Evidence from the adoption of value-added measures*. Unpublished manuscript, UC Riverside.

Borg, M. G., & Riding, R. J. (1991). Occupational stress and satisfaction in teaching. *British Educational Research Journal*, *17*(3), 263–281.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2011). Teacher layoffs: An empirical illustration of seniority v. measures of effectiveness. *Education Finance and Policy*, *6*(3), 439–454.

Browne, M. W. (2010). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150.

Chapman, D. W., & Lowther, M. A. (1982). Teachers' satisfaction with teaching. *Journal of Educational Research*, *75*(4), 241–247.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–2679.

Chute, E. (2013, August 13). New teacher evaluation process set to begin in Pittsburgh Public Schools. *Pittsburgh Post-Gazette*.

Condie, S., Lefgren, L., & Sims, D. (2014). Teacher heterogeneity, value-added and education policy. *Economics of Education Review*, *40*(1), 76–92.

Cullen, J. B., Jacob, B. A., & Levitt, S. D. (2005). The impact of school choice on student outcomes: An analysis of the Chicago public schools. *Journal of Public Economics, 89*(5/6), 729–760.

Dee, T., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267–297.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2014). Choosing the right growth measure: Methods should compare similar schools and teachers. *Education Next*, *14*(2), 66–71.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272.

Fitzpatrick, M. D., Grissmer, D., & Hastedt, S. (2011). What a difference a day makes: Estimating daily learning gains during kindergarten and first grade using a natural experiment. *Economics of Education Review*, *30*(2), 269–279.

Fowler, F. J. (2013). *Survey research methods*. Washington, DC: Sage.

Goldhaber, D., Cowan, J., & Walch, J. (2013). Is a good elementary teacher always good? Assessing teacher performance estimates across subjects. *Economics of Education Review*, *36*(1), 216–228.

Goldhaber, D., Gross, B., & Player, D. (2011). Teacher career paths, teacher quality, and persistence in the classroom: Are public schools keeping their best? *Journal of Public Policy and Management*, *30*(1), 57–87.

Goldhaber, D., & Theobald, R. (2013). Managing the teacher workforce in austere times: The determinants and implications of teacher layoffs. *Education Finance and Policy*, *8*(4), 494–527.

Grissom, J. A., Nicholson-Crotty, S., & Harrington, J. R. (2014). Estimating the effects of No Child Left Behind on teachers' work environments and job attitudes. *Educational Evaluation and Policy Analysis*, *36*(4), 417–436.

Guarino, C. M., Santibanez, L., & Daley, G. (2006). Teacher recruitment and retention: A review of the recent empirical literature. *Journal of Educational Research*, *76*(2), 173–208.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201–209.

Hanushek, E. A. (2009). Teacher deselection. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession* (pp. 165–180). Washington, DC: Urban Institute.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*(2), 267–271.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*(3), 393–416.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635.

Ingersoll, R. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal*, *38*(3), 499–534.

Jepsen, C., Mueser, P., & Troske, K. (2010). *Labor-market returns to the GED using regression discontinuity analysis* (Working Paper No. 10-14). Columbia, MO: University of Missouri Department of Economics.

Johnson, S. M. (2006). *The workplace matters: Teacher quality, retention and effectiveness*. Washington, DC: National Education Association.

Johnson, S. M., Kraft, M. A., & Papay, J. P. (2012). How context matters in high-need schools: The effects of teacher working conditions on their professional satisfaction and their students' achievement. *Teachers College Record*, *114*(10), 1–39.

Kraft, M. A. (2015). Teacher layoffs, teacher quality and student achievement: Evidence from a discretionary layoff policy. *Education Finance and Policy*, *11*(4), 1–41.

Kraft, M. A., & Gilmour, A. F. (2016). *Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness* (Working paper). Brown University, Providence, RI.

Lawler, E. E., III, & Porter, L. W. (1967). The effect of performance on job satisfaction. *Industrial Relations*, *7*(1), 20–28.

Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, *142*(2), 655–674.

Liu, X. S., & Ramsey, J. (2008). Teachers' job satisfaction: Analyses of the teacher follow-up survey in the United States for 2000–2001. *Teaching and Teacher Education*, *24*(5), 1173–1184.

Lorenzo-Seva, U. (2013). *How to report the percentage of explained common variance in exploratory factor analysis* (Technical report). Department of Psychology, Universitat Rovira i Virgili, Tarragona.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698–714.

McCrary, J., & Royer, H. (2011). The effect of female education on fertility and infant health: Evidence from school entry policies using exact date of birth. *American Economic Review*, *101*(1), 158–195.

McClellan, C. A., & Donoghue, J. R. (2013). *Evaluation of Tennessee's Teacher Evaluation Systems*. Nashville, TN: Tennessee Consortium on Research, Evaluation, and Development.

The New Teacher Project. (2012). *The irreplaceables: Understanding the real retention crisis in America's urban schools*. New York City, NY: Author.

Papay, J. P., Murnane, R. J., & Willet, J. B. (in press). The impact of test-score labels on human-capital investment decisions. *Journal of Human Resources*.

Rockoff, J. E., Staiger, D. O., Kane, T. J., & Taylor, E. S. (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review*, *102*(7), 3184–3213.

Rothstein, J. (2015). Teacher quality policy when supply matters. *American Economic Review*, *105*(1), 100–130.

Schafer, J. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, *8*, 3–15.

Schwab, D. P., & Cummings, L. L. (1970). Theories of performance and satisfaction: A review. *Industrial Relations*, *9*(4), 408–430.

Shifrer, D., Turley, R. L., & Heard, H. (2013). *Houston Independent School District's Aspire Program: Estimated effects of receiving financial awards*. Houston, TX: Houston Educational Research Consortium Policy Report.

Springer, M. G., Swain, W. A., & Rodriguez, L. A. (2016). Effective teacher retention bonuses: Evidence from Tennessee. *Educational Evaluation and Policy Analysis*, *38*, 199–221.

Steele, J., Pepper, M., Springer, M. G., & Lockwood, J. R. (2015). The distribution and mobility of effective teachers. *Economics of Education Review*, *48*, 86–101.

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB Era. *Education Finance and Policy*, *11*(3), 340–359.

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy*, *10*(4), 535–572.

Stockard, J., & Lehman, M. B. (2004). Influences on the satisfaction and retention of 1st-year teachers: The importance of effective school management. *Educational Administration Quarterly*, *40*(5),742–771.

Strunk, K. O., Weinstein, T. L., & Makkonnen, R. (2013). *Sorting out the signal: Do multiple measures of teachers' effectiveness provide consistent information to teachers and principals?* (Working paper). University of Southern California.

Stutz, T. (2014, August 3). Texas teachers, principals ignore survey ordered by legislature. *Dallas News*. Retrieved from http://www.dallasnews.com/news/education/headlines/20140803-texas-teachers-principals-ignore-survey-ordered-by-legislature.ece

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, *102*(7), 3628–3651.

Tennessee State Board of Education. (2004). *Framework for evaluation & professional growth: Comprehensive assessment*. Retrieved from http://www.tn.gov/education/teaching/framework/comprehensive_assessment.pdf

United States Department of Education. (2009). *Race to the Top*. Washington, DC: Author.

United States Department of Education. (2014). *Race to the Top*. Washington, DC: Author. Retrieved from http://www.whitehouse.gov/issues/education/k-12/race-to-the-top.

United States Government Accountability Office. (2013). *Race to the Top: States implementing teacher and principal evaluation systems despite challenges*. Washington, DC: Author.

Van Der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review*, *43*(4), 1249–1287.

Watts, M., & Becker, W. E. (2008). A little more than chalk and talk: Results from a third national survey of teaching methods in undergraduate economics courses. *The Journal of Economic Education*, *39*(3), 273–286.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New York, NY: The New Teacher Project.

Winters, M. A., & Cowen, J. M. (2013). Would a value-added system of retention improve the distribution of teacher quality? A simulation of alternative policies. *Journal of Policy Analysis and Management*, *32*(3), 634–654.

Manuscript received October 26, 2015
Final revision received October 21, 2016
Accepted November 4, 2016